HMMA401 Statistique Computationnelle 2020/2021
Jean-Michel Marin

Vous devez envoyer **avant le jeudi 18 février 2021 à 17h00** un rapport écrit de 15 pages maximum (hors annexes) **au format pdf** à l'adresse suivante `jean-michel.marin@umontpellier.fr`.

Le rapport doit clairement indiquer quelle a été la contribution de chaque membre du projet. Le listing du code R utilisé doit figurer en annexe du rapport.

Projet 4 Bayesian inference for logit and probit models

The **bank** dataset is made of four measurements on 100 genuine Swiss banknotes and 100 counterfeit ones (file `bank.txt`). The response variable y is thus the status of the banknote, where 0 stands for genuine and 1 stands for counterfeit, while the explanatory factors are the length of the bill x_1, the width of the left edge x_2, the width of the right edge x_3, and the bottom margin width x_4, all expressed in millimeters.
We want a probabilistic model that predicts the type of banknote (i.e., that detects counterfeit banknotes) based on the four measurements above.
We use the notation $\mathbf{y} = (y_1, \ldots, y_n)$ for a sample of n responses and

$$\mathbf{X} = [\mathbf{x}_1 \quad \ldots \quad \mathbf{x}_k] = \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1k} \\ x_{21} & x_{22} & \ldots & x_{2k} \\ x_{31} & x_{32} & \ldots & x_{3k} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix}$$

for the $n \times k$ matrix of corresponding explanatory variables, possibly with $x_{11} = \ldots = x_{n1} = 1$. We use y and $\mathbf{x}$ as generic notations for single-response and covariate vectors, respectively. Once again, we will omit the dependence on $\mathbf{x}$ or $\mathbf{X}$ to simplify notations.
A *generalized linear model* is specified by two functions:

(i) a conditional density f of y given $\mathbf{x}$ that belongs to an exponential family and that is parameterized by an expectation parameter $\mu = \mu(\mathbf{x}) = \mathbb{E}[y|\mathbf{x}]$ and possibly a dispersion parameter $\phi > 0$ that does not depend on $\mathbf{x}$; and

(ii) a *link* function g that relates the mean $\mu = \mu(\mathbf{x})$ of f and the covariate vector, $\mathbf{x}$, as $g(\mu) = (\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta})$, $\boldsymbol{\beta} \in \mathbb{R}^k$.

For identifiability reasons, the link function g is a one-to-one function and we have

$$\mathbb{E}[y|\boldsymbol{\beta}, \phi] = g^{-1}\left(\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta}\right) .$$

We can thus write the (conditional) likelihood as

$$\ell(\boldsymbol{\beta}, \phi|\mathbf{y}) = \prod_{i=1}^{n} f\left(y_i|\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta}, \phi\right)$$

if we choose to reparameterize f with the transform $g(\mu_i)$ of its mean and if we denote by $\mathbf{x}^i$ the covariate vector for the ith observation.
The most widely used GLMs are presumably those that analyze binary data, as in **bank**, that is, when $y_i \sim \mathcal{B}(1, p_i)$ (with $\mu_i = p_i = p(\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta})$). The mean function p thus transforms a real value into a value between 0 and 1, and a possible choice of link function is the *logit transform*,

$$g(p) = \log\{p/(1-p)\}\,,$$

associated with the *logistic regression model*.
Because of the limited support of the responses y_i, there is no dispersion parameter in this model and the corresponding likelihood function is

$$\begin{aligned}\ell(\boldsymbol{\beta}|\mathbf{y}) &= \prod_{i=1}^{n}\left(\frac{\exp(\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta})}{1+\exp(\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta})}\right)^{y_i}\left(\frac{1}{1+\exp(\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta})}\right)^{1-y_i} \\ &= \exp\left\{\sum_{i=1}^{n} y_i\,\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta}\right\}\Big/\prod_{i=1}^{n}\left[1+\exp(\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta})\right]\,.\end{aligned} \tag{1}$$

It thus fails to factorize conveniently because of the denominator: there is no manageable conjugate prior for this model, called the *logit model*.
There exists a specific form of link function for each exponential family which is called the *canonical link*. This canonical function is chosen as the function $g^\star$ of the expectation parameter that appears in the exponent of the natural exponential family representation of the probability density, namely

$$g^\star(\mu) = \theta \quad \text{if} \quad f(y|\mu,\varphi) = h(y)\exp\varphi\{T(y)\cdot\theta - \Psi(\theta)\}\,.$$

Since the logistic regression model can be written as

$$f(y_i|p_i) = \exp\left\{y_i\,\log\left(\frac{p_i}{1-p_i}\right) + \log(1-p_i)\right\}\,,$$

the logit link function is the canonical version for the Bernoulli model. Note that, while it is customary to use the canonical link, there is no compelling reason to do so, besides following custom!
For binary response variables, many link functions can be substituted for the logit link function. For instance, the *probit* link function, $g(\mu_i) = \Phi^{-1}(\mu_i)$, where Φ is the standard normal cdf, is often used in econometrics.

$$\ell(\boldsymbol{\beta}|\mathbf{y}) \propto \prod_{i=1}^{n}\Phi(\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta})^{y_i}\,\left[1-\Phi(\mathbf{x}^{i\mathrm{T}}\boldsymbol{\beta})\right]^{1-y_i}\,. \tag{2}$$

We use the following hierarchical prior distribution

$$\boldsymbol{\beta}|\sigma^2 \sim \mathcal{N}_k\left(\mathbf{0}_k, \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\right) \quad \text{and} \quad \pi(\sigma^2) \propto \sigma^{-2}.$$

1 Give the density of the posterior distributions (up to normalizing contants) for the logit (1) and the probit (2) models. We get non-standard densities. That is our target densities.

2 Use an Hastings-Metropolis algorithm based on a multivariate Gaussian random walk to get some realizations coming approximately from the two target posterior distributions.

As starting value, one can use the maximum likelihood estimation $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ and as scale matrix the estimated covariance matrix $\hat{\Sigma}$ of the maximum likelihood estimator.

```
mod <- summary(glm(y~-1+X,family=...))
```

provides $\hat{\boldsymbol{\beta}}$ in mod$coeff[,1] and $\hat{\Sigma}$ in mod$cov.unscaled.

4 Compare the maximum likelihood and the two bayesian estimates of $\boldsymbol{\beta}$.